Large-Sample Estimation

Outline

- Statistical Inference
- Types of Estimators
- Point Estimation
- Interval Estimation
- Estimating the Difference between Two Population Means
- Estimating the Difference between Two Population Proportions
- Choosing the Sample Size

Statistical Inference

- **Statistical inference** – making decisions or predictions about *parameters* of a *population*, e.g. μ, σ, and binomial proportion p.

- Two major categories of making inferences
  - Estimation – predicting the value of the parameters
  - Hypothesis testing – making a decision about the value of a parameter based on some perception of what the value might be.

- The **goodness of the inference** evaluates the accuracy of the method used in doing statistical inferences.

Types of Estimator

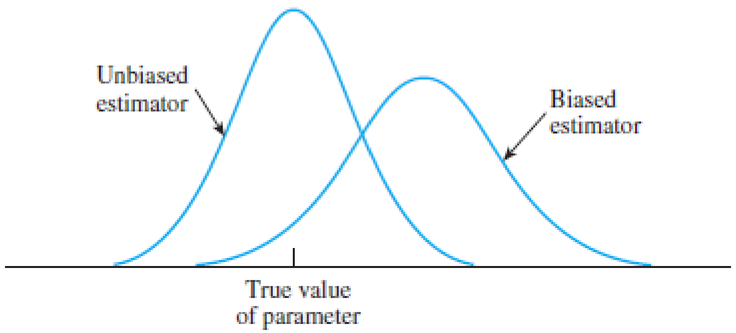
- **Estimator** – a rule, usually expressed as a formula, to calculate an estimate based on information in a sample.
  - **Point estimator**: rule or formula to calculate the estimate of *a population parameter*. The resulting value is called **point estimate**.
  - **Interval estimator**: rule or formula to calculate *two values* defining the interval within which the parameter is expected to be. The resulting values are called **interval estimate** or **confidence interval**.

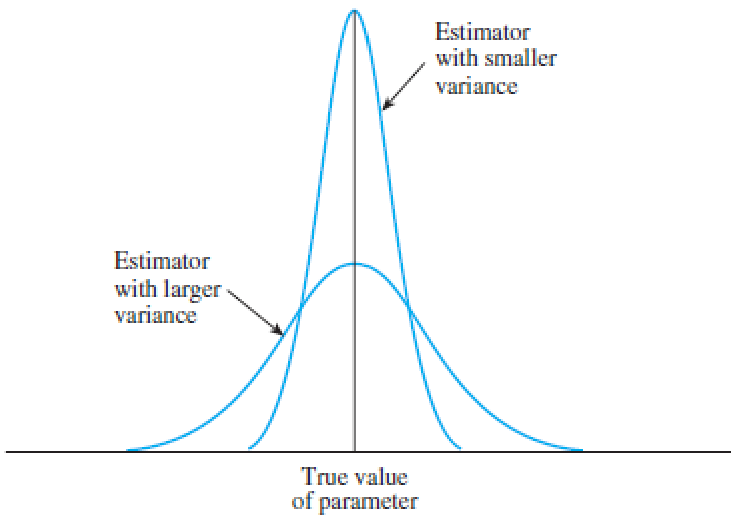
Point Estimation

- Several sample statistics can be used as an estimate for a population parameter. Best point estimator should satisfy the following

    - Sampling distribution of a point estimator should be **unbiased**, i.e. the *mean* of the distribution equals the true value of the parameter.

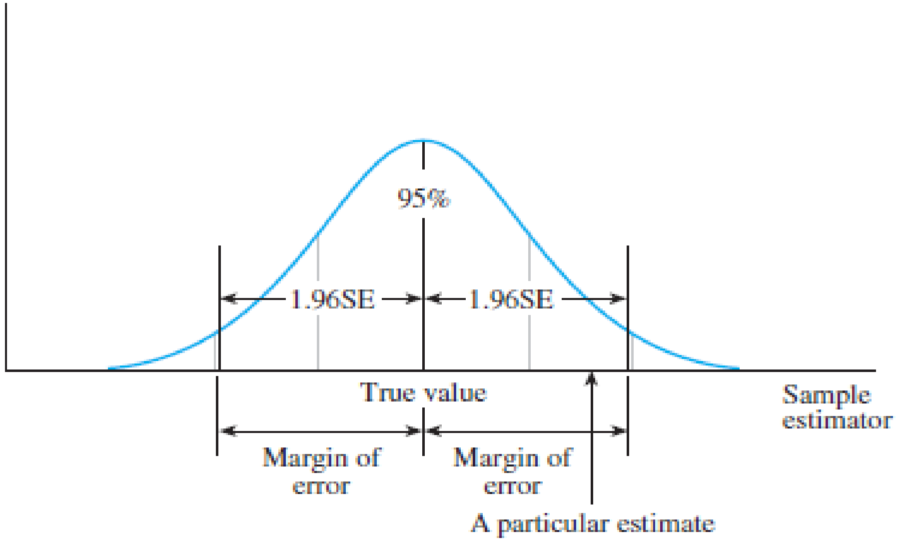
    - The *spread* of the sampling distribution should be as small as possible, i.e. the resulting estimate are more likely to be near the true value of the parameter.

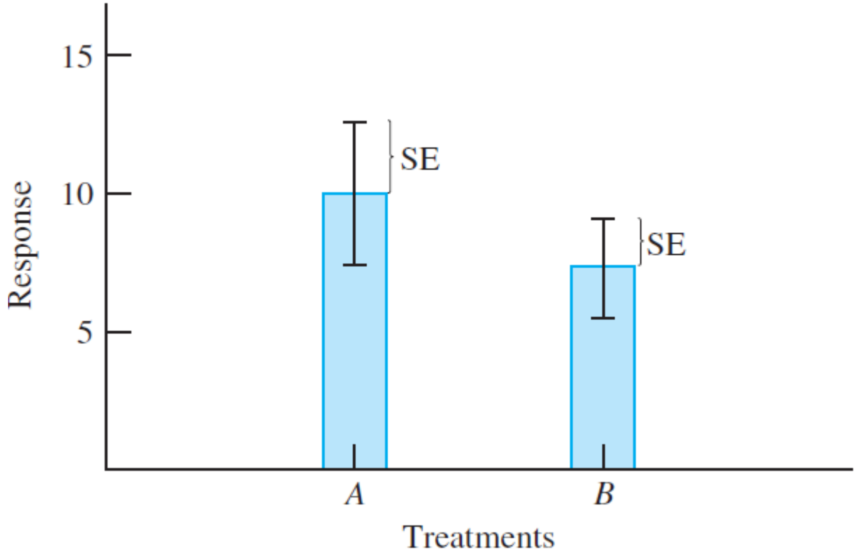




Point Estimation of $\boldsymbol{\mu}$ and binomial proportion $\boldsymbol{p}$

- We can reasonably assume that the sample sizes are *large* => the sampling distribution is a *normal distribution* and *centers* around the parameter being estimated (thanks to CLT)
- Therefore 95% of all points estimated will be within 1.96 standard deviation around the mean – called the 95% **margin of error**.
- Note that the *estimated standard error* (from the sampling process) is usually considered a *reasonable approximate* of the *true standard error* of the population.
- Point estimation is normally reported together with either standard deviation s or the standard error $\frac{s}{\sqrt{n}}$






THUAN AN
Rượu
Bàn Than
Dung tích: 500ml
Độ cồn: 29%+-2

Point Estimation

Example. A random sampling of 200 savings account in a local community showed an average increase in savings account values of 7.2% over twelve months with a standard deviation of 5.6%. What is the mean percent increase of savings account values over the last 12 months for the whole community?

The point estimate of mean percent increase of savings values for the whole community μ is $\bar{x}$=7.2%. The 95% margin of error of this estimation can be approximated by using the sample standard deviation, which is

$$\pm 1.96 \frac{s}{\sqrt{n}} = \pm 1.96 \frac{5.6}{\sqrt{200}} = \pm .776\%$$

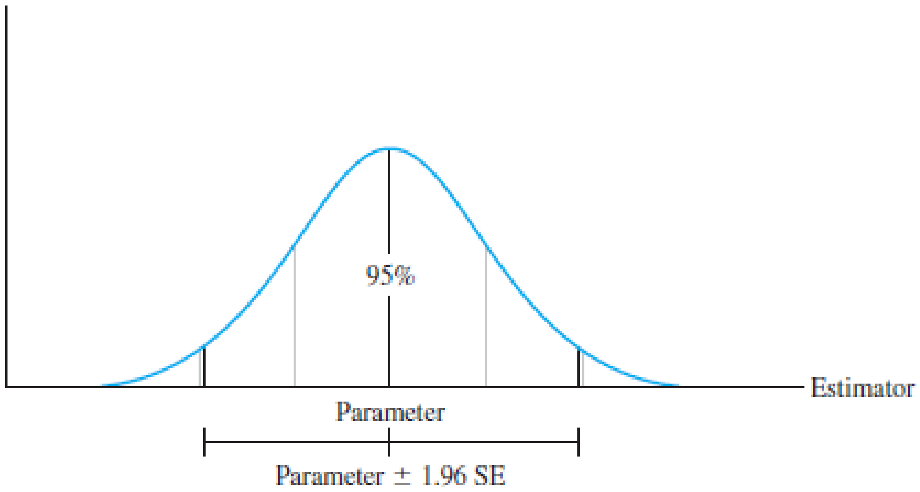
Example. Interviewing 900 registered voters in the US showed that 51% of them believed that US should drop the number of legal immigrants. The sample proportion $\hat{p}$=.51 is the best point estimate for the proportion of all registered voters who believed that the number of legal immigrants should be dropped. The margin of error can also be approximated by using $\hat{p}$, $1.96\mathrm{SE} = 1.96\sqrt{.51 * .94/900} = 0.0167$

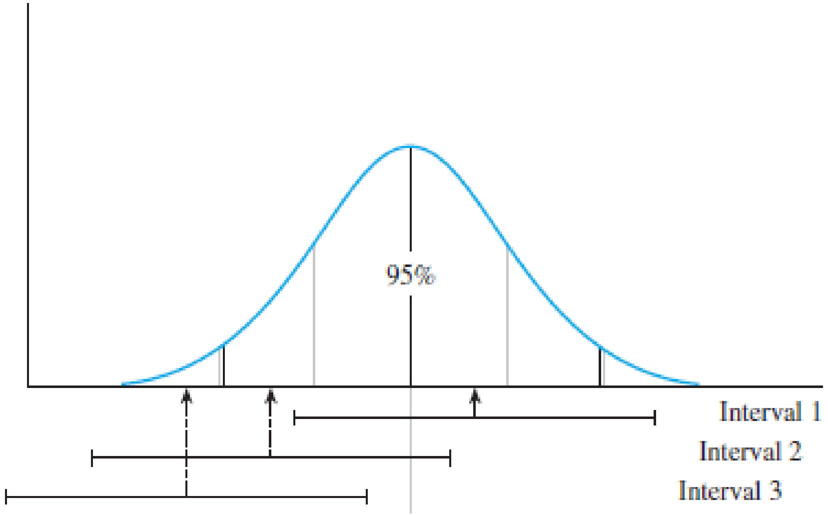
Interval Estimation

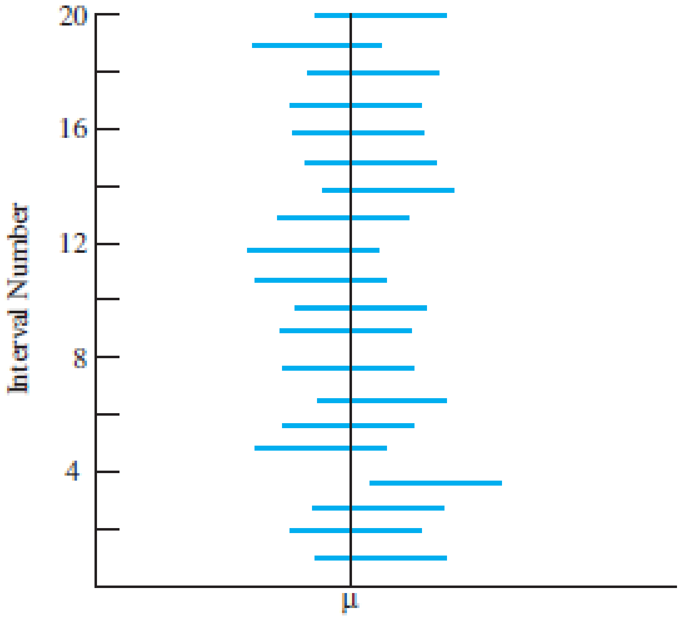
- Interval estimator – rule or formula to calculate 2 values of an interval that there is “a high probability” to contain the parameter of interest.
- **Confidence coefficient $(\mathbf{1-\alpha})$** – the probability that a confidence interval will contain the estimated parameter
- For example, 95% confidence interval means the probability that the interval will contain the estimated value is 95%

Constructing a confidence interval

- For an approximately normal sampling distribution, a 95% confidence interval is calculated by $parameter \pm 1.96SE$, and is *approximated* by $estimator \pm 1.96SE$. (Note that SE of the population ≈ SE from the sample)
- How often does this interval include the parameter being estimated? 95% of the repeated samples will contain your parameter, e.g. μ.



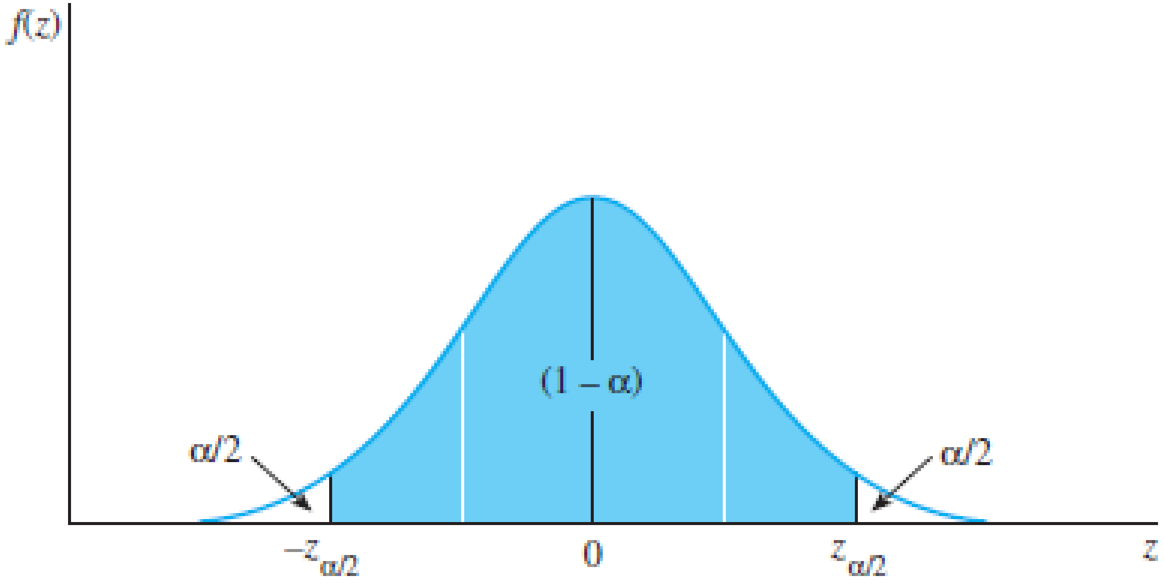




Constructing a confidence interval

- The above calculating law can be applied to other confidence coefficient $1-\alpha$, by the below general formula

$$(point\ estimate) \pm z_{\frac{\alpha}{2}}(standard\ error\ of\ the\ estimator)$$



Confidence coefficient, $(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.98	.02	.01	2.33
.99	.01	.005	2.58

Confidence Interval for Population Mean

- According to CLT when sample size is large, $\bar{x} \approx \mu$, and $s \approx \sigma$, the (1- α)% confidence interval calculated by

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

- Example: A random sample of 500 vehicles registered was selected, 68 of which were SUV. What is the 95% confidence interval to estimate the proportion of SUV in the population of registered vehicles?

Confidence Interval for Population Proportion

- According to CLT when sample size is large, the sample proportion $\hat{p}$ is the best point estimate for the population proportion p.
- Because the sampling distribution of $\hat{p}$ is approximately normal, with mean p and standard error $\sqrt{p(1-p)/n}$, the (1- α)% confidence interval is

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}$$

- Example: A survey of 1002 adults showed that 39% were against abortion. What is 90% confidence interval for the proportion of adult Americans being against abortion?

The sample proportion $\hat{p} = 39\%$ can be used as point estimate for p. Because $\hat{p}$ is approximately normal distributed, the 90% confidence interval around $\hat{p} = 39\%$ is

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} = .39 \pm 1.645\sqrt{.39 * .61/1002} = .39 \pm .025 = [.365, .415]$$

Estimating Difference between 2 Population Means

- Given 2 sets of population and an independent random sample drawn from each of the populations as summarized below

	Population 1	Population 2
Mean	μ_1	μ_2
Variance	σ_1^2	σ_2^2

	Sample 1	Sample 2
Mean	$\bar{x}_1$	$\bar{x}_2$
Variance	s_1^2	s_2^2
Sample Size	n_1	n_2

- The difference between two sample means should provide information on the actual difference between the two population means.

Estimating Difference between 2 Population Means

- The sampling distribution of $\overline{x_1} - \overline{x_2}$ follow a normal distribution
  - If the populations are normally distributed, or
  - If the populations are not normally distributed and the sample size is large.

- The mean of the sampling distribution of $\overline{x_1} - \overline{x_2}$ is $\mu_1 - \mu_2$

- The standard error of the distribution is $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (the approximation applies when the sample size is large)

Estimating Difference between 2 Population Means

- **Point estimation** of $\mu_1 - \mu_2$
  - Because $\mu_1 - \mu_2$ is the mean of the sampling distribution of $\overline{x_1} - \overline{x_2}$, $\overline{x_1} - \overline{x_2}$ is the unbiased point estimator of $\mu_1 - \mu_2$.
  - The 95% margin of error is $\pm 1.96SE = \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- $(1-\alpha)$% **confidence interval** for $\mu_1 - \mu_2$

$$(\overline{x_1} - \overline{x_2}) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Confidence interval is usually the *preferred* estimation for estimating the difference between 2 population means.

Estimating Difference between 2 Population Means

- Example: Sample values for daily intakes of dairy products

	Men	Women
Sample Size	50	50
Sample Mean	756	762
Sample Standard Deviation	35	30

- Point estimate of $\mu_1 - \mu_2$ = $\overline{x_1} - \overline{x_2} = 756 - 762 = -6$
- The standard error $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{50^2}{50} + \frac{30^2}{50}} = 6.52$
- 95% confidence interval $-6 \pm 1.96 * 6.52$=-6 ± 12.78
- Because $-18.78 < \mu_1 - \mu_2 < 6.78$, we should NOT conclude there is a difference between the 2 populations.

Estimating Difference between 2 Population Proportions

- Assuming independent random samples having n_1 and n_2 trials are drawn from two binomial populations 1 and 2 that have parameters p_1 and p_2, respectively.
- The unbiased estimator of $(p_1 - p_2)$ is $(\widehat{p_1} - \widehat{p_2})$
- The mean of the sampling distribution of $(\widehat{p_1} - \widehat{p_2})$ is $(p_1 - p_2)$
- The standard error of the sampling distribution of $(\widehat{p_1} - \widehat{p_2})$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

- $\widehat{p_1}$ and $\widehat{p_2}$ should be approximately normal, i.e. $n_1\widehat{p_1}$, $n_1(1-\widehat{p_1})$, $n_2\widehat{p_2}$, $n_2(1-\widehat{p_2})$ should all be larger than 5.

Estimating Difference between 2 Population Proportions

- **Point estimation** of $(p_1 - p_2)$ is $(\widehat{p_1} - \widehat{p_2})$, with 95% margin of error

$$\pm 1.96SE = \pm 1.96\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

- A $(1-\alpha)$% **confidence interval** for $(p_1 - p_2)$ is

$$(\widehat{p_1} - \widehat{p_2}) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

- $\widehat{p_1}$ and $\widehat{p_2}$ should be approximately normal, i.e. $n_1\widehat{p_1}$, $n_1(1-\widehat{p_1})$, $n_2\widehat{p_2}$, $n_2(1-\widehat{p_2})$ should all be larger than 5.

Estimating Difference between 2 Population Proportions

Example: Voting for bond proposal for school construction, residents in the developing section vs the rest of the city. What is the difference between the true proportions favoring the proposal with 99% confidence interval?

	Developing Section	Rest of the City
Sample Size	50	100
Number Favoring Proposal	38	65
Proportion Favoring Proposal	.76	.65

The 99% confidence interval is calculated as

$$(\widehat{p_1} - \widehat{p_2}) \pm z_{0.005}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} = (.76 - .65) \pm 2.58\sqrt{\frac{.76 * .24}{50} + \frac{.65 * .35}{100}}$$

The resulting 99% confidence interval is (-.089,.309), which means that we should NOT conclude that the proportion favoring the proposal are different.

Choosing the Sample Size

- Sample size is crucial to reliability and goodness of inferences made by researchers – through the measurements of *margin of error* and the *width of the confidence interval*.
- The sample size can be determined based on the requirement for margin of error B, and the desired confidence coefficient $(1-\alpha)$, represented by $z_{\alpha/2}$

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < B \text{ or } n > \left(z_{\alpha/2}\frac{\sigma}{B}\right)^2$$

- The population standard deviation σ is usually unknown and can be estimated by standard deviation of previous samples or, if not available, $\sigma \approx Range/4$.

Choosing the Sample Size

Parameter	Estimator	Sample Size	Assumptions
μ	$\bar{x}$	$n \geq \dfrac{z_{\alpha/2}^2 \sigma^2}{B^2}$	
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$n \geq \dfrac{z_{\alpha/2}^2(\sigma_1^2 + \sigma_2^2)}{B^2}$	$n_1 = n_2 = n$
p	$\hat{p}$	$n \geq \dfrac{z_{\alpha/2}^2 pq}{B^2}$	
		or	
		$n \geq \dfrac{(.25)z_{\alpha/2}^2}{B^2}$	$p = .5$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$n \geq \dfrac{z_{\alpha/2}^2(p_1q_1 + p_2q_2)}{B^2}$	$n_1 = n_2 = n$
		or	
		$n \geq \dfrac{2(.25)z_{\alpha/2}^2}{B^2}$	$n_1 = n_2 = n$ and $p_1 = p_2 = .5$